Part 1: Simulation Exercise

Austin L. Bistline

January 4, 2017

Overview

In this project I investigate the exponential distribution in R and compare it with the Central Limit Theorem which states that the distribution of averages is often normal, even for very abnormal distributions such as the exponential distribution. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda λ is the rate parameter. The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. $\lambda = 0.2$ for all of the simulations. I investigate the distribution of averages of 40 exponentials, performing 1000 simulations.

I illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials by:

1. Showing the sample mean and comparing it to the theoretical mean of the distribution.
2. Showing how variable the sample is (via variance) and comparing it to the theoretical variance of the distribution.
3. Showing that the distribution is approximately normal.

Generating the Sample Means of the Exponential Distributions

Sample means are generated by first creating a 40x1000 matrix of exponential distributions(using rexp() as described in the Overview), then using the mean() function for 40 exponentials across each of the 1000 columns. This is accomplished with the replicate() function.

```
rm(list = ls()) # Remove previous variables for consistency
set.seed(123) # For reproducibility
library(ggplot2) # Import libraries
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
lambda = 0.2
num_sims = 1000
num_edists = 40
edistributions = replicate(num_sims, rexp(num_edists, lambda), simplify = "matrix")
dim(edistributions)
```

```
## [1]   40 1000
```

The mean of 40 exponentials is calculated from 1000 exponential distributions using the apply() function, which takes as arguments a matrix, 1 or 2 to specify rows or columns (respectively), and the function to apply (mean() in this case). The output should be an array of 1000 values, and this is checked using the length() function.

```
edist_means = apply(edistributions, 2, mean)

# Should be an array with 1000 indexes
length(edist_means)
```

```
## [1] 1000
```

Sample Mean

The mean of 1000 averages is calculated and is not only found to be very close to the theoretical mean of 5, but found to converge on 5 as more and more samples are included.

```
# Sample Mean
mean(edist_means)
```

```
## [1] 5.011911
```

```
# Theoretical Mean
1/lambda
```

```
## [1] 5
```

The difference between the empirical and theoretical means is minimal as seen below.

```
# Difference between Sample Mean and Theoretical
1/lambda - mean(edist_means)
```

```
## [1] -0.01191128
```

In a graphical comparison, the mean converges on 5 as more exponential distribution means are considered in the average (Figure 1).

```
means = cumsum(edist_means)/(1:num_sims)
```

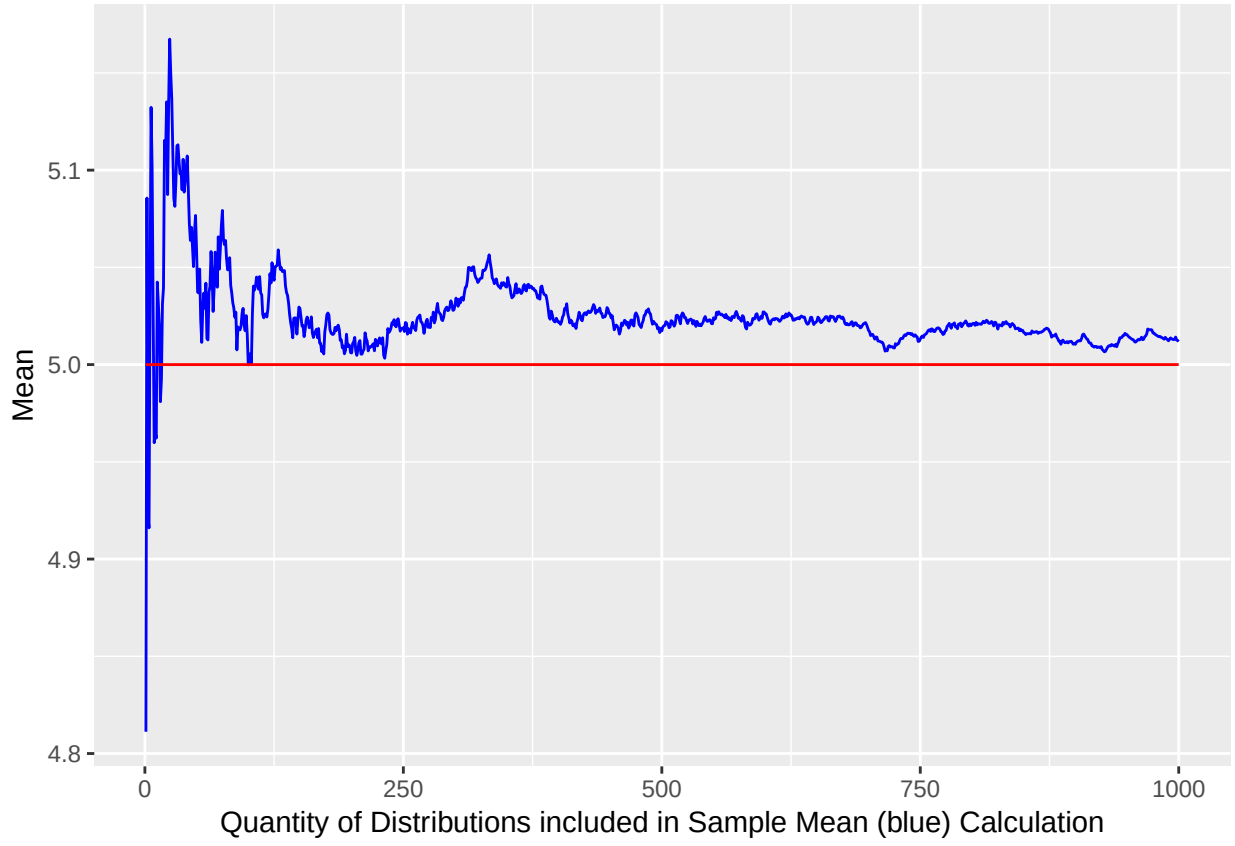

Figure 1: A comparison of the empirical (blue) vs theoretical (red) mean of the exponential distributions. As more distributions are included in the calculation, the sample mean converges on the theoretical mean.

Variance of the distribution of means

Theoretical variance is $1/(\lambda^2 n)$, where n is the number of means in the distributions, and the expression in this case evaluates to 0.625. The sample variance comes close to this value as can be observed in Figure 2.

```
# Sample Variance
var(edist_means)
```

```
## [1] 0.6004928
```

```
# Theoretical Variance
1/(lambda^2)/num_edists
```

```
## [1] 0.625
```

```
variances = as.vector((length(edist_means) - 1))
for (i in 2:length(edist_means)) {variances[i-1] = var(edist_means[1:i])}
```

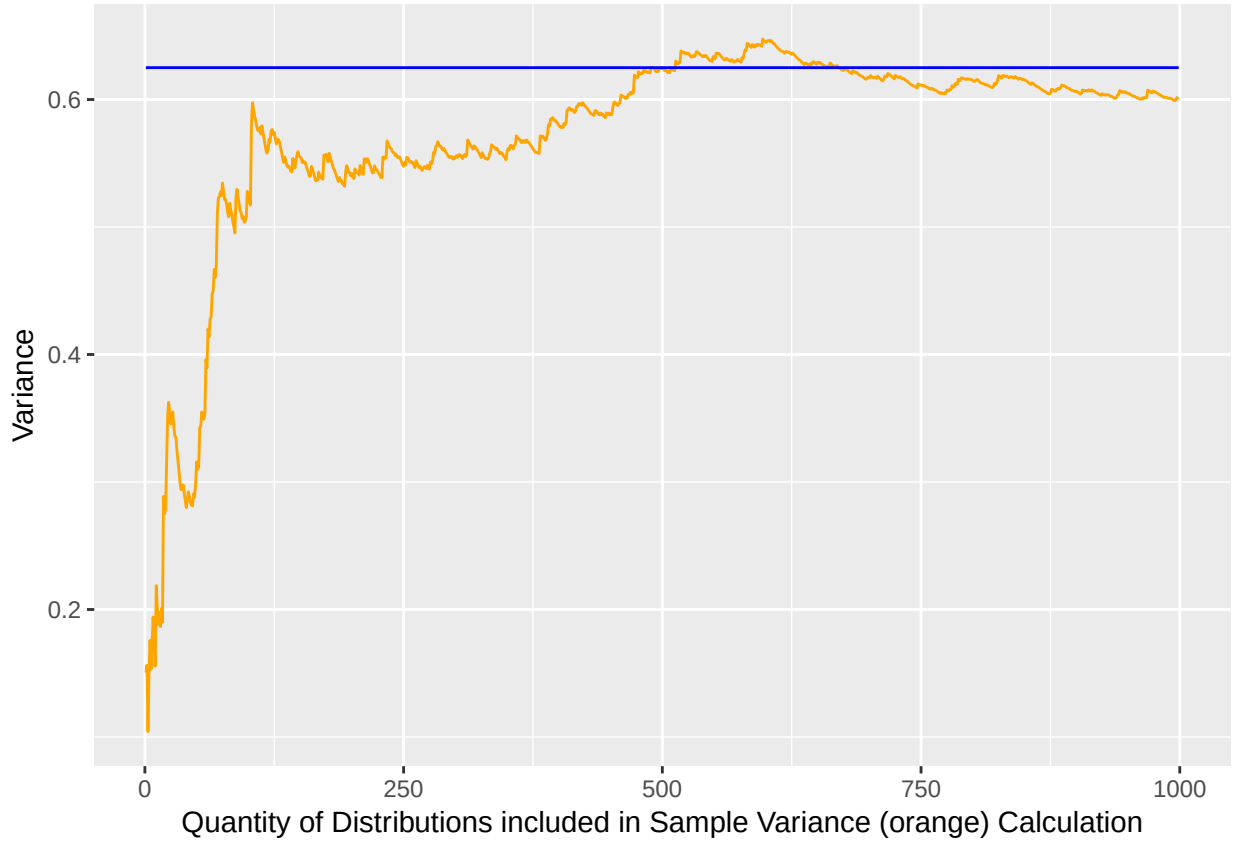

Figure 2: A comparison of the sample (orange) vs theoretical (red) variance of the exponential distributions. As more distributions are included in the calculation, the sample variance converges on the theoretical variance.

Approximately-Normal Distribution

A histogram of the exponential-average distribution is similar to a normal distribution of the same μ and variance, and is therefore approximately normal (Figure 3). A Normal Q-Q plot of the exponential-average distribution in Figure 4 indicates that the quantiles are approximately linear as well. A more normal distribution produces a Normal Q-Q plot with more linearity as can be seen in Figure 5.

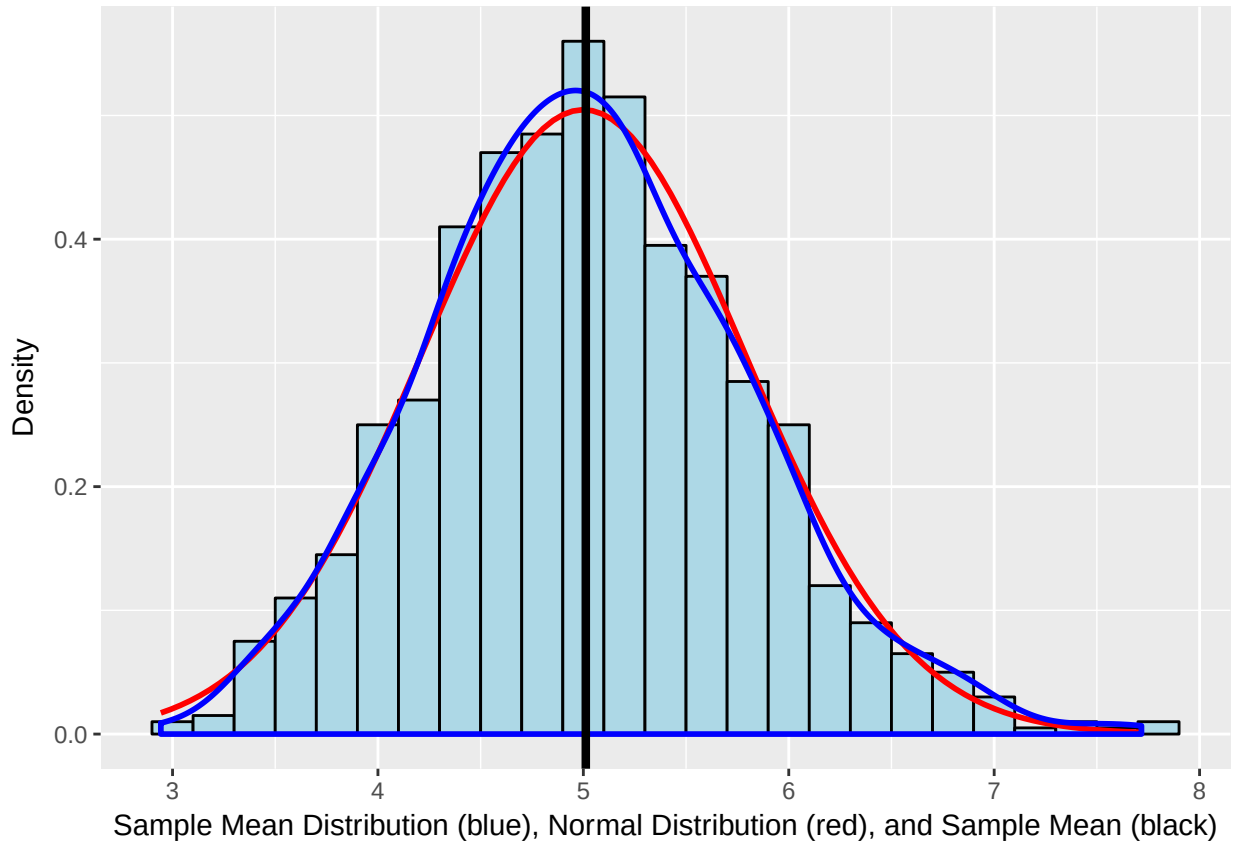

Figure 3: The histogram of exponential distribution means (light blue with dark-blue) outline compared to a normal distribution (red). The exonential distribution means have an approximately-normal distribution. Note that the black vertical line marks the sample mean.

```
qqnorm(edist_means)
qqline(edist_means, col = 2)
```

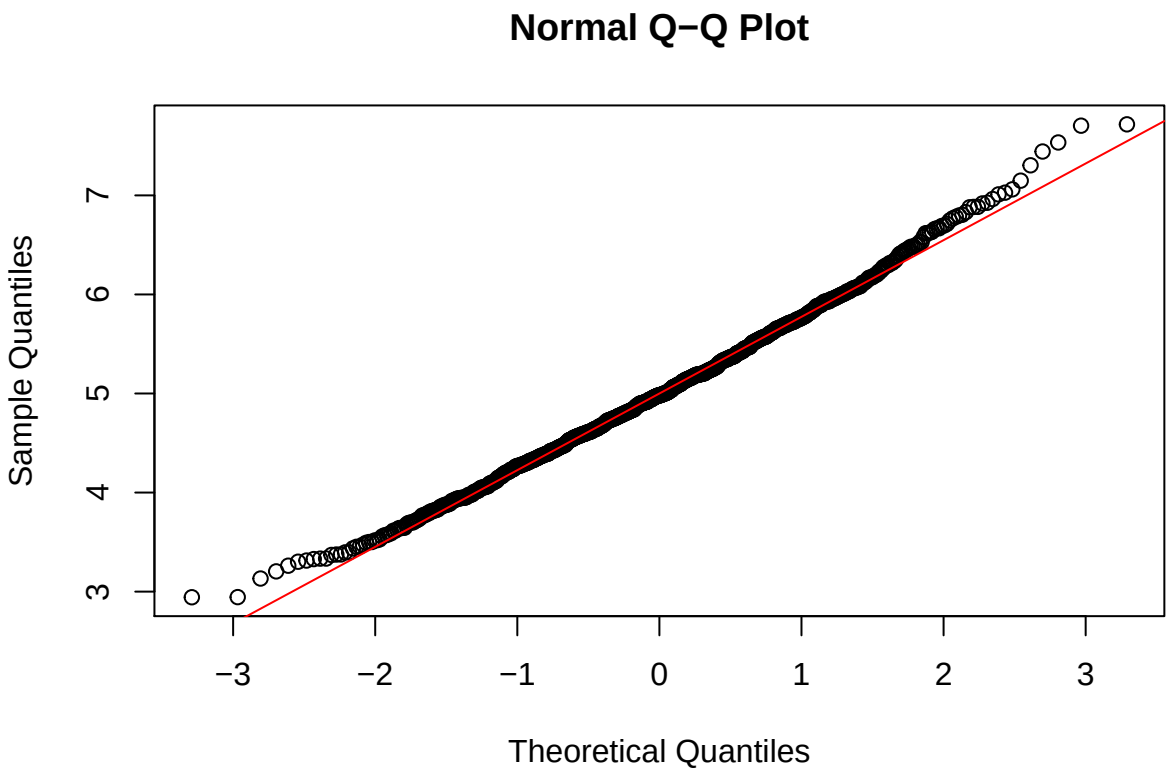

Figure 4: The Normal Q-Q plot of the exponential distribution means (black). The means do not exactly follow the Q-Q plot of a precisely normal distribution (red) which is linear, but are close and could be considered normally distributed when compared to a perfect uniform distribution as seen in Figure 5.

```
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(runif(40) * 10))
qqnorm(mns)
qqline(mns, col = 2)
```

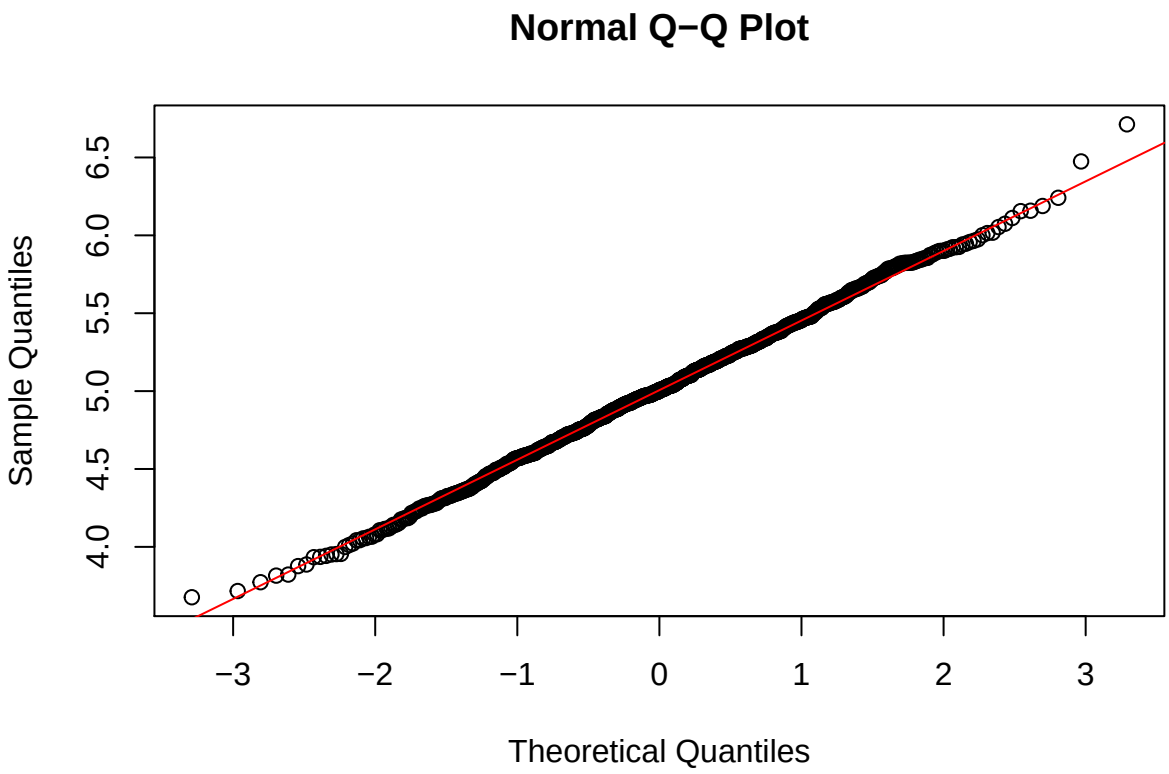

Figure 5: The Normal Q-Q plot of 1000 averges of 40 random uniform values (above0 is more linear than the similar mean of 40 exponentials from 1000 exponential distributions seen in Figure 5.